CCDH

Center for Countering Digital Hate

Don’t Feed the Trolls

How to Deal with Hate on Social Media

Contents

Introduction 3

Troll Propagandists 4

The Troll Counter-Strategy 9

Recommendations for the Media 10

Recommendations for Social Media companies 11

Selected Bibliography 12

Acknowledgements 12

CCDH
Center for Countering Digital Hate

The Center for Countering Digital Hate deals with the increasing use of racial and religious intolerance, sexism, homophobia, and other forms of identity-based hate to polarise societies and undermine democracy. We find practical solutions to this threat, producing policy solutions and actions we can all take to counter hate.

Hateful discourse hardens differences, coarsens the public sphere and leads to anger and violence.

Proponents of hate are quick to adopt new techniques and technologies, particularly social media, to spread their ideas and beliefs.

We've seen it in the UK and US, throughout Europe, and further afield – from Brazil to Nigeria, India to Israel.

Populist politicians have allied with haters to gain new advocates and troll armies which harass opponents and institutions that provide checks and balances in liberal states, like the media and judiciary. If we do not deal with this threat, it will continue to divide our societies and undermine the strength of our democratic systems.

CCDH works with practitioners in diverse fields, such as political science, behavioural science, the law, countering violent extremism and counterterrorism, child protection and identity-based hate to develop strategies that strengthen tolerance, liberalism and democracy, and counterstrategies to new forms of political hate.

"I learned long ago, never to wrestle with a pig. You get dirty, and besides, the pig likes it."
George Bernard Shaw

Introduction

Don't Feed the Trolls is a practical guide for public figures on how best to deal with hate-filled online trolls. Over recent months we have been asked to advise on dealing with hurtful troll "storms". As there is currently no clear guidance, we reviewed evidence from diverse fields.

What we quickly realised is that targeted abuse of public figures is not just about "lulz"; it is a deliberate tactic designed to generate outrage and attract more viewers to misinformation and hate. Political trolls are skilled and determined propagandists, promulgating harmful extremist beliefs, like sectarianism, racism and religious intolerance, using abuse and mockery. The specific ideologies advocated by trolls vary considerably. We have not tried to define trolls too tightly as adhering to a particular ideology or approach. Rather, we point to the use of *ad hominem* attacks, sectarianism and deliberate offensiveness as key characteristics.

Troll propagandists know that when public figures engage with abusive trolls to defend their reputation or their values – a normal behavioural response – this inadvertently spreads and legitimates the trolls' message and tricks social media algorithms into pushing it into a broader array of users' timelines.

This helps trolls to spread their message to far more people than they could alone, as they usually congregate in small, densely- interconnected, highly-active clusters with little access to the broader public. Social media has given trolls a cost-free means to spread their message to far larger numbers of people. This has exposed more people to hateful attitudes and conspiracism, and risks the trolls winning new converts.

The accelerating spread of hate-filled trolls' beliefs shows that we have not fully comprehended how to respond to the complex webs of interests dynamically aligning around political figures and movements. Only by fully understanding and exposing their tactics can we respond effectively.

To aid public figures, we have developed a harm-minimisation framework for dealing with trolling and not becoming part of the problem. We aim to limit the corrosive impact of troll tactics on public discourse and use evidence-based approaches and tools to protect ourselves as individuals and as a society.

This report does not talk about state-sponsored propaganda or automated technologies like bots. They are almost certainly involved in tactically amplifying this activity, but this is not our focus here. Nor is this report specifically for active campaigners, who choose to take part in political debates. This is for public figures who do not wish to amplify hate, and yet have inadvertently become crucial vectors in its spread.

Imran Ahmed
Chief Executive Officer

Dr Linda Papadopoulos
Chair of the Advisory Board

Troll Propagandists

The stereotype of a troll - a young person taking pleasure in sending malicious abuse from a laptop in their bedroom - is, at best, a half-truth. Trolling has become an anarchic, subversive, postmodern subculture, with its origins on websites such as 4Chan, 8chan and Reddit. As Whitney Phillips explains in *This is Why We Can't Have Nice Things,* trolls self-define as people who believe "nothing should be taken seriously" and "like to disrupt stupid conversations on the Internet" for "lulz": "amusement derived from another person's anger."

But trolls competing for lulz can't influence politics alone. They need weaponizing by savvy propagandists, who help to coalesce trolls and conventional political actors around issues of mutual interest, so they might campaign to achieve their common political aims.

Extremist ideologues know that identity-based hate is rejected by the general public in liberal societies and so they subsume its tropes and themes under politically-charged issues where the conversation seems safer. Issues of common agreement between trolls and populist politicians, such as a revulsion for "political correctness" or "corrupt" traditional elites, is a way to start discussions that resonate with wider swathes of the public but can then be used by bad actors to inject hate.

Social media is essential to this process. By spreading complex ideas using simple visual memes, and locating, targeting and demoralising opponents with abuse, propagandists can both shape public opinion and bombard any resistance until they retreat.

These tactics work. We know they do. We see them every day on social media, an environment that has become an important way for many people to find information and learn what others think.

Trolling tactics have been adopted around the world by movements seeking to change the knowledge, opinions and values of our societies – in short, our culture – to help achieve their political goals. The methods they use to shape culture include more than just trolling of course. Fake news sites preach misinformation, promulgating "alternative facts" that seek to help justify normally unattractive political philosophies. Social media spaces, such as Facebook groups and WhatsApp groups, are potential platforms for radicalising by extremists. Trolls on Twitter are just one component, albeit one of the most visible, by which culture – our shared knowledge, norms and social mores – is being reformed.

It is crucial to remember that troll propagandists are not seeking to enter an honest debate. They and their active troops are not there to be converted. Trolls motivated by hate want to spread their aberrant worldview and precursor beliefs designed to be difficult to disprove. For example, Islamophobic trolls claim that all Muslims are potential extremists, and that a core tenet of Islam is to lie to cover up one's true beliefs and intentions[1], thus any protestations of innocence are to be distrusted or, worse, examples of their fundamental mendacity. Antisemitic trolls claim that Jews control the media, and simultaneously that the disavowal of

[1] Aka the "taqiyya" conspiracy theory.

this fraudulent idea by journalists or academics armed with "facts" is proof that Jews do indeed control the media.

It is a basic behavioural response to wish to correct someone we believe is wrong. But that instinctual response must be overcome because engaging with erroneous material and allegations legitimises that views and lends it credibility it does not deserve.

There is a swathe of research in political psychology, the study of radicalisation and epistemology that shows the difficulties in trying to dissuade someone who is wedded to a worldview may, in some cases, even entrench that view.[2] The mere act of repeating a claim to refute it can entrench it. As one study explains: "attempting to quash rumors through direct refutation may facilitate their diffusion by increasing fluency... merely repeating a rumor increases its power."[3]

We need instead to stop legitimating troll propaganda by engaging with them and raising their salience in the overall debate – thus increasing both the exposure people have to hate-filled trolls' arguments and their ability to persuade.

Any solution must centre on minimising harm by reducing exposure and the spread of misinformation and lies that underlies hate.

Steve Bannon depicted in the style of Pepe the Frog, a popular Alt-Right meme. Posted on Reddit by an anonymous poster.

[22] Brendan Nyhan, Jason Reifler "*When Corrections Fail: The Persistence of Political Misperceptions*," (Political Behavior, 32, 303 – 333). It is also worth reading Porter, Ethan, Wood, Thomas J. "*The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence*,"(Political Behavior. 41 (1): 135-163).

[3] Berinsky, Adam, "*Rumors and Health Care Reform: Experiments in Political Misinformation*," (B.J.Pol.S. 47: 241-262).

The Troll Playbook

"We should always be on the lookout for any opportunity to grab media attention... The most obvious way to do this is to troll public figures and get them to whine about it. I keep thinking this will stop working eventually, but it just never does."

Trolling playbook, published by a White Nationalist Group [4]

If we were sat having a quiet conversation in a crowded room, we would most likely go unnoticed. If we started shouting at each other, everyone would know we were there and what we were talking about. Observers might pick a side. Some would choose based on the merits of the argument. Others based on identity or because one side was, say, being belligerent or abusive. Even if one argument was demonstrably wrong it is almost inevitable that at least one person would pick that side. This is the simple dynamic that trolls use to spread hate when they target public figures – abuse causes outrage, which attracts attention, driving awareness, consideration and conversion.

For the main part, society has fallen for the trolls' tactics. We act counter-productively, engaging trolls, debating them, believing this is a battle of ideas. In fact, the trolls are playing a quite different game. They don't want to "win" or "lose" an argument; they just want their ideas to be heard by as many potential converts as possible.

Trolls Enjoy Hurting People

When trolls abuse public figures, it is quite often intentionally hurtful; like a bully, they know where to hit to make it hurt the most. Psychologists Naomi Craker and Evita March studied the personality characteristics that drive trolling behaviour and found that trolls actually gain pleasure from seeing someone else hurt; something they call "negative social potency".[5] As Dr March explains, "while antisocial personality traits do play a role, what really influences trolling behaviour is the social pleasure derived from knowing that others are annoyed by it. The more negative social impact the troll has, the more their behaviour is reinforced."[6]

In short, reacting to trolls really *is* giving them exactly what they want. Furthermore, using the same logic, the language of victimisation, even if used to highlight the issue, often acts as positive reinforcement for trolls.

[4] *Daily Stormer Style Guide*, accessed June 2019. This document lays out in detail the techniques used by modern Nazi digital propagandists.

[5] Craker, Naomi & March, Evita. (2016). "*The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours*." Personality and Individual Differences. 102. pp79-84.

[6] March, Evita. Evita March: *"Don't feed the trolls' really is good advice - here's the evidence."* New Zealand Herald. 10th October 2016. Retrieved from: https://www.nzherald.co.nz/technology/news/article.cfm?c_id=5&objectid=11725624

"We have an army at our disposal, to use for whatever purposes we wish. They enjoy attacking people on the internet. Organizing campaigns is good in every way: it energizes our people, gives them something fun to do, gains media attention and increases [our] level of general infamy."

Trolling playbook, published by a White Nationalist Group [7]

Defending Yourself is a Trap

Trolls exploit the natural human instinct of self-defence. Social media is a new medium where our natural instincts can sometimes cause us to fall into traps and do harm without realising it.

When the targets of abuse by trolls respond they do three things. First, they rebroadcast troll memes to their own followers. Second, they legitimise a dangerous ideology as valid for discussion – an honest point of view that you might disagree with, but nevertheless deserves the merit of acknowledgement and discussion. Third, they confirm to the trolls that their targets are listening to them and are affected by what they see online, which reinforces the trolls' behaviour.

Propagandists know that all you have to do to spread a conspiracy theory is to get more people to hear it. The problem for the target of their activity is that the formats of most social media platforms allow little substantial debate. Complex arguments require more than the 280 characters to explain that Twitter allows, for example. Nor does it show the relative merits of a point of view. To the casual observer a thread looks like a debate between two different sides of an argument – the form legitimises the content regardless of the merits of either side.

Studies by University of Cambridge Conspiracy and Democracy Project with YouGov found that over half of American and British adults believe in at least one common conspiracy theory.[8] While the precise numbers are disputed, the research reminds us that untrue beliefs, based on aberrant information, are widespread. Unpicking a conspiracy theory often requires persuading people long- or deeply held beliefs are untrue, which is remarkably difficult to do. These beliefs can be deeply entrenched, protected by selectively cherrypicked information that backs up their erroneous belief, leading to a deep and abiding faith in something that is in fact demonstrably untrue.

By inadvertently helping to expose their followers to conspiracy theories and aberrant information by engaging with trolls, public figures almost inevitably expand the number of people believing a rumour or conspiracy. Trolls know this; they do not seek to convert those who actively disagree with them; they seek to consolidate their hold over those who might already hold some prior misinformation or to misinform those

[7] *Daily Stormer Style Guide*

[8] https://www.cam.ac.uk/research/news/brexit-and-trump-voters-more-likely-to-believe-in-conspiracy-theories-survey-study-shows

who hold no strong beliefs or have no knowledge about the topic at all. Quite often the techniques used by trolls – for example using memes to convey false information quickly and in a format that maximises the probability of information retention – are simply more effective than the techniques used by those trying to refute their arguments.

Engagement with trolls ultimately can cause considerable harm by giving them a much wider audience than they could achieve on their own and, ultimately, helps them grow their support base. Indeed, analysis by the social media consultancy, Signify, carried out exclusively for this report, showed a series of incidents in which precisely this pattern emerged.[9]

Trolls Attack as a Pack

Troll propagandists coordinate activities to "isolate" and "mass-assault" a public figure as a core tactic.[10] At its simplest, this involves identifying a target and an issue that might induce a reaction, but might include other elements, for example, an agreed hashtag to maximise the likelihood that disparate abuse by numerous actors "trends" as a single topic. Trolls are encouraged to target "journalists, bloggers, politicians and public figures" as well as "entertainment figures."[11]

A target might be bombarded with dozens, even hundreds of messages in a short space of time. It is understandable if a hundred Twitter accounts with, say, a thousand followers each, tweet at someone, the target might assume that all the followers of those people have seen the attack. In this example, a hundred troll accounts with one thousand followers each might mean up to one hundred thousand people (100 x 1000) have seen it. But, in reality, that isn't the case.

Troll networks are small, coordinated, densely interconnected and highly active clusters. What this means is that trolls within a cluster are usually following each other and therefore each trolls' followers are for the main part within the same troll network. The total audience for one hundred trolls whose followers are within the same troll network of a thousand trolls accounts is really not much more that network itself. However, an unwitting target can be fooled into thinking that they are being exposed to much broader humiliation and ridicule. This is a vital tactic for fringe groups because it vastly amplifies the threat they *appear* to pose.

The reality, however, is that they are small groups of fringe believers who have little influence on the broader public without the undeserved and inadvertent amplification caused by engaging with them.

[9] For more information on Signify, please visit www.signify.ai
[10] Daily Stormer Style Guide
[11] ibid

The Troll Counterstrategy

Things to Do Immediately

Trolls want you to engage with their content to raise visibility of their propaganda. Our aim is, instead, to reduce harm to you and others.

Resist the urge to respond. You might think winning or losing relies on whether you "win the argument". But for a troll, winning or losing is about how many people see their propaganda. They don't have to convert many of your followers to increase their numbers substantially.

When a troll targets you for abuse, **block them immediately**; this will ensure that they cannot tweet at you ever again, and removes mentions of them from your notifications. It is remarkable how few people you actually have to block to stop a troll storm in its tracks. Furthermore, it will stop their small networks from being able to target you again in future.

If you receive several tweets in a short period of time, temporarily **switch off app notifications on your mobile devices**; this will protect you from unplanned exposure to troll hate.

Do not post saying that you are being targeted; this will simply invite further abuse and sympathy, all of which raises troll content up in prominence.

A troll storm is unpleasant. Even after blocking abuse and switching off notifications, you will still have to deal with having been abused in the first place. It is important to **get space from social media and show yourself some compassion**. They attack you because they are trying to exploit your achievements and the number of normal people who take an interest in what you do and say on social media.

Things to Do Next

Once you have dealt with the initial rush of posts, or in the event that a storm starts to gain a life of its own, for example, if the trolls include "blue tick" verified users, you should consider using the following additional tactics. These require more time but can be vital in the battle against trolling tactics.

Record

If you feel a message you have received is defamatory or might contain criminal content, e.g. incitement of violence or harassment; or content that glorifies terror, then record it. The best way to do this is to take a screenshot. Make sure to capture the time, date and sender.

Report

Posts can be reported to the social media company directly from the interface. Do so; these platforms will often prioritise complaints coming from public figures. If the content is potentially criminal, then keep a record and seek advice on whether and how to report it to the police or to discuss it with a defamation lawyer. If you need advice get in touch with an expert.

Recruit Help

Finally, you are not on your own. There are organisations with experience and expertise in online hate, social media tactics and trolling who are there to help.

You can contact anti-hate organisations directly or, without referencing the trolls' propaganda, share anti-hate material as a way of ensuring your followers are exposed to arguments for tolerance.

We are also happy to provide advice. You can contact us at info@counterhate.co.uk or on Twitter at @CCDHate.

Recommendations for the Media

Journalists and media organisations are a strategic focus of troll delegitimation.

The lessons for public figures hold strong here. We know that journalists will have the same instinct to defend themselves, their hard-earned reputations and their values as anyone else. Helping journalists to recognise the unrepresentative nature of influences that might affect their cognition and behaviour is doubly important given their important role in informing the public.

However, the media has also served to amplify the tactics of trolls. This is not a new debate. Many British Muslims held their hands up in dismay and frustration at the singular focus on Islamists when they were featured on the news regularly for deliberately provocative and highly offensive stunts. Similarly, the disproportionate focus given to the loudest voices, no matter how unrepresentative they might be, in the 2016 election and today to the purveyors of hate online explains, in part, their success.

Trolls seek to have their voices amplified through the media. And the best way to do that is to say something controversial. When Carl Benjamin (aka Sarkon of Arkkad) casually discussed raping Jess Phillips, his comments were of course disgusting, but they were then amplified to such an extent that MPs and commentators were forced to respond. This then created a debate on whether it was a joke or not.

Trolling is often a coordinated effort, with the aim of influencing journalists' professional output. By engaging, journalists inadvertently send a message that they're listening, which just gives them succour.

Our recommendations to the media are:

- Do not amplify hate by filing easy but pointless reports on hashtag trends targeting an individual, especially where they do not reflect public opinion but rather are a stunt by trolls.
- Avoid treating troll narratives as worthy. Plucking tweets out of the Twittersphere to back up an argument that there was "outrage" about a public figure is easy enough, but it can amplify the worst kinds of voices and ultimately compounds the harm done to innocents.
- Work with mental health professionals on a new code to deal with bullying online where you do report on it.

Recommendations for Social Media companies

Social media companies run on a core business model where the core metric is “time spent” online. As such they benefit financially from anything that boosts activity. This, unfortunately, includes storms that start as a result of bullying or harassing behaviour.

Algorithms need to be rewritten to integrate language analysis that actively and effectively deprioritises abusive language that might indicate bullying or harassment. The existing solutions have simply not worked effectively.

When social media companies publish and therefore promote “trends” they act as publishers, not simply as platforms.

The publishing of trends centred on bullying individuals serves to highlight and intensify both the bullying activity and raises prominence and awareness of troll propaganda. Social media companies are rightly under huge scrutiny for the important part they play in modern society.

We recommend they do the following:

- Social media companies should engage in productive collaboration with anti-hate and mental health experts to fundamentally rethink how they adjust their systems to deal with existing and emerging troll tactics; social media companies owe an enormous duty to the societies whose collective wealth and endeavour they profit from and have a duty of care to that society.
- Introduce human moderation of the publishing of “trends” to exclude anything that comprises unreasonable harassment or bullying behaviour.
- Make it easier for people to identify when they are at the centre of a storm and work with users to develop a mode that works across devices that reduces notifications to the bare minimum, causes human moderators to scrutinise the activity and take proactive action against bullies.
- Include richer data in notifications on who is communicating with you. For example, identify potential bot accounts or high-activity, low-influence trolls by including data on posts per day and number of followers in the notification itself.

Selected Bibliography

Haidt, Jonathan, "*The Righteous Mind: Why Good People are Divided by Politics and Religion*," (Penguin Books, 2012)

Lakoff, George, "*Don't Think of an Elephant! Know Your Values and Frame The Debate: The Essential Guide for Progressives*" (Chelsea Green Publishing, 2004)

Mudde, Cas and Kaltwasser, Rovira, "*Populism: A Very Short Introduction*," (Oxford University Press, 2017)

Nagle, Angela, "*Kill All Normies: Online Culture Wars from 4Chan and Tumblr to Trump and the Alt-Right*" (John Hunt Publishing, 2017)

Oliver, J. Eric and Wood, Thomas J, "*Conspiracy Theories and the Paranoid Style(s) of Mass Opinion*," (American Journal of Political Science 58 (4): 952–66)

Phillips, Whitney, "*This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*," (The MIT Press, 2015)

Rosenblum, Nancy L and Muirhead, Russell, "*A Lot of People Are Saying: The New Conspiracism and the Assault on Democracy*," (Princeton University Press, 2019)

Schmitt, Carl, "*The Concept of the Political,"* (University of Chicago Press, 1996)

Sunstein, Cass R, "*On Rumours: How Falsehoods Spread, Why We Believe Them, What Can be Done*," (Penguin Books, 2011)

Sunstein, Cass R, "*Going to Extremes: How Like Minds Unite and Divide*," (Oxford University Press, 2009)

Acknowledgements

The Center wishes to thank the following colleagues who provided valuable insight or feedback in the writing of this report. The report does not claim to represent their views but their assistance was of value:

Dr Daniel Allington, Kings College London
Simon Clark, Center for American Progress
Dr Kate Ferguson, Protection Approaches
Dr Rob Ford, University of Manchester
Dr Siobhain McAndrew, University of Bristol
Morgan McSweeney
Euan Neill, Commission for Countering Extremism
Hannah O'Rourke
Randeep Ramesh
Rachel Riley
Dave Rich, Community & Security Trust
Jonathan Sebire, Signify
Will Somerville, Unbound Philanthropy

© 2019 Center for Countering Digital Hate Ltd